Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 4 (2025)

Impact of Deep Learning and Deepfakes in E-Governance: An Indian Government Perspective

Mr. Pawan Kadyan , Dr. Madhu Bala Royand Dr. Abhishek Roy

Inspector General of Registration &Commissioner of Stamp Revenue, Government of West Bengal, India. E-mail: pawan.kadyan@gmail.com, pawan.kadyan@ias.nic.in
Associate Professor, Department of Management, Techno India University, Salt Lake, Kolkata, West Bengal, India. Email: drmadhubalaroy@gmail.com
Chief Technology Officer (CTO, Department of Personnel& Administrative Reforms, Government of West Bengal, India. Email: drroyabhishek@gmail.com

Abstract

With the advancement of the artificial intelligence (AI) usage in all the aspects of business applications, technology developments, predictions and cost reduction it has become very easy to use AI for profit generation, delivering services and cost reduction. It is also being used in many e-governance programs to give the citizen services without much hassle and cost. It is helping decision making much faster and data driven in governance. However, with every technology advancement comes its darker side. Here the darker side of AI is that it can be used for fraudulent activities, distorted information, guised social engineering attacks, trapping, fake arrests etc. To curb this menace its highly important to build program, policies and platforms tamper proof and robust.

Keywords: Deep fake, deep learning, deepfakes, e-governance, policy framework, AI

Introduction

Recent advances in generative deep learning—Generative Adversarial Networks (GANs) and diffusion models—have made it inexpensive and fast to produce synthetic audio, video, and images indistinguishable from genuine media for many consumers. When combined with targeted dissemination through social media platforms and instant messaging, deepfakes pose direct risks to democratic processes, public administration, and trust in digital government services. India has already faced notable incidents and responded with advisories and regulatory activity; CERT-In and MeitY have published advisories and guidance aimed at curbing deepfake harms. CERT-In+1

This paper frames the deepfake threat specifically against Indian e-governance by (a) mapping attack vectors and targets, (b) providing projections for incident trajectories under alternative mitigation strategies, (c) quantifying impacts on citizen trust, and (d) proposing a prioritized policy and technical roadmap.

1. Background and Related Work

1.1 Technical foundations: Deep learning and synthetic media

Modern deepfakes are generated using GANs, autoencoders, and diffusion-based models. These architectures can synthesize photorealistic faces and highly convincing voice clones from small datasets. Detection methods range from feature-based detectors to deep-learning classifiers trained on manipulated media datasets; however, attacker and defender advances are in an arms race, and generalization across unseen synthesis methods remains challenging. Recent surveys and methodological overviews provide comprehensive taxonomies of detection approaches. MDPI+1

1.2 Policy landscape in India

India's regulatory posture has evolved from non-binding advisories to stronger operational expectations for platforms. CERT-In issued a high-severity advisory on deepfakes in November 2024, and MeitY has released guidance and draft rule amendments to address synthetically generated information and platform responsibilities. The government has also issued advisories around elections and information integrity, and various states have reported FIRs and police action in response to viral deepfake incidents. CERT-In+2Press Information Bureau+2

1.3 Documented incidents and social impact

High-visibility incidents — including fabricated videos and audio of public figures and corporate executives — show how deepfakes can influence public debate and financial markets. Coverage during election cycles and regulatory reactions indicate a material level of harm already realized in India. WIRED+1

2. Threat Model and E-Governance Attack Surfaces

2.1 Actors and capabilities

- **Adversaries:** State-affiliated actors, political operatives, financially motivated fraudsters, malicious insiders, and opportunistic pranksters.

- **Capabilities:** Access to public footage, facial images, and voice samples; use of cloud compute and open-source generative models; distribution via social platforms and private messaging.

2.2 Targets in Indian e-governance

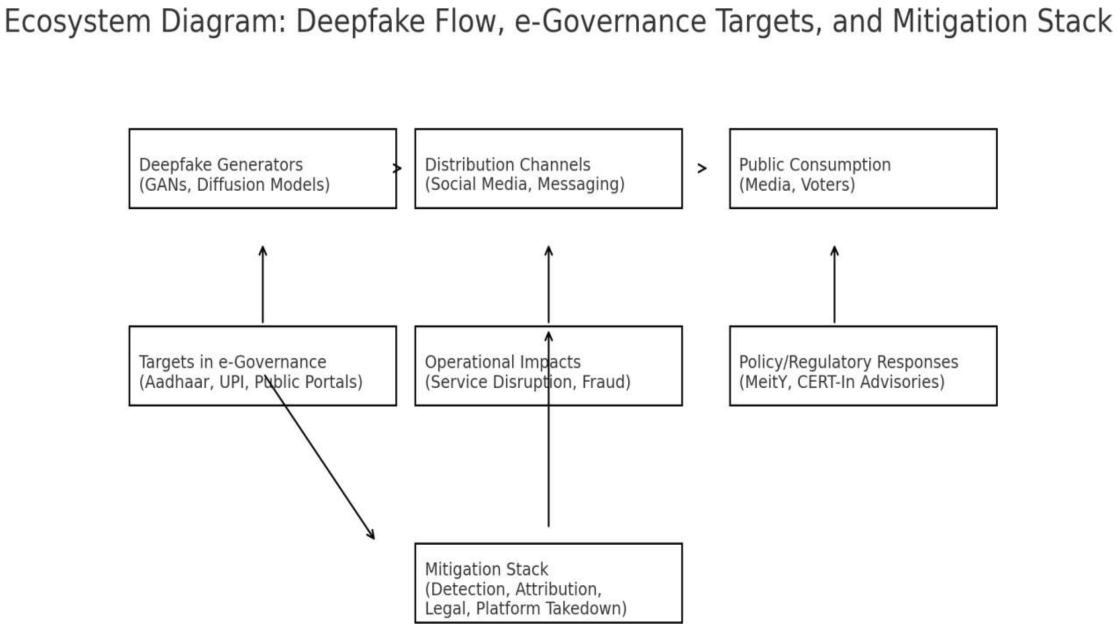
- **Identity systems:** Aadhaar enrollment and authentication pathways—risks include deepfake-based identity fraud and impersonation.
- **Payment systems:** UPI and banking interactions may be targeted through synthetic voice/video to authorize or social-engineer payments.
- **Public communication channels:** Government announcements, politician messaging, and voter outreach that can be spoofed to manipulate opinion.
- **Service Portals:** Manipulated media intended to defame, extort, or cause reputational harm to officials and departments.

2.3 Attack pathways

1. **Creation:** Generate synthetic media using small datasets.
2. **Amplification:** Disseminate via social media, messaging apps, and coordinated bots.
3. **Consumption:** Public interprets manipulated media as authentic, potentially causing rapid misinformation spread.
4. **Operational impact:** Fraud, service disruption, erosion of trust, and administrative

overhead.

A schematic of this flow is provided in the ecosystem diagram (see Figure 1).


Ecosystem Diagram: Deepfake Flow, e-Governance Targets, and Mitigation Stack
Deepfake Generators (GANs, Diffusion Models)
Distribution Channels (Social Media, Messaging)
Public Consumption (Media, Voters)
Targets in e-Governance (Aadhaar, UPI, Public Portals)
Operational Impacts (Service Disruption, Fraud)
Policy/Regulatory Responses (MeitY, CERT-In Advisories)
Mitigation Stack (Detection, Attribution, Legal, Platform Takedown)

(Authorsframework)

3. Methodology: Projections and Trust Modeling

3.1 Scenario definitions

We model three mitigation scenarios over a 10-year horizon (2023–2032):

- **Slow mitigation:** Limited detection capacity and weak platform takedown; incident counts grow with the baseline generative trend.
- **Moderate mitigation:** Deployment of detection tools at scale, platform takedown commitments, public-awareness campaigns; incident growth slows and plateaus.
- **Aggressive mitigation:** Mandatory detection infrastructure for critical e- governance services, rapid takedown, mandatory provenance/watermarking, legal enforcement; incidents decline substantially.

3.2 Projection model (simple, transparent)

Let I_base(t) be the baseline estimated incident count without mitigation (derived from historical reports and trend extrapolation). We apply scenario multipliers m_s(t) to produce I_s(t) = I_base(t) × m_s(t). Parameters were set to reflect conservative to aggressive reductions for illustrative policy analysis. (Numerical values are hypothetical and meant for planning rather than precise prediction.)

3.3 Citizen Trust Model

We model citizen trust U as an index in [0,1] influenced by observed incident rates and mitigation investment M (index 0–100):

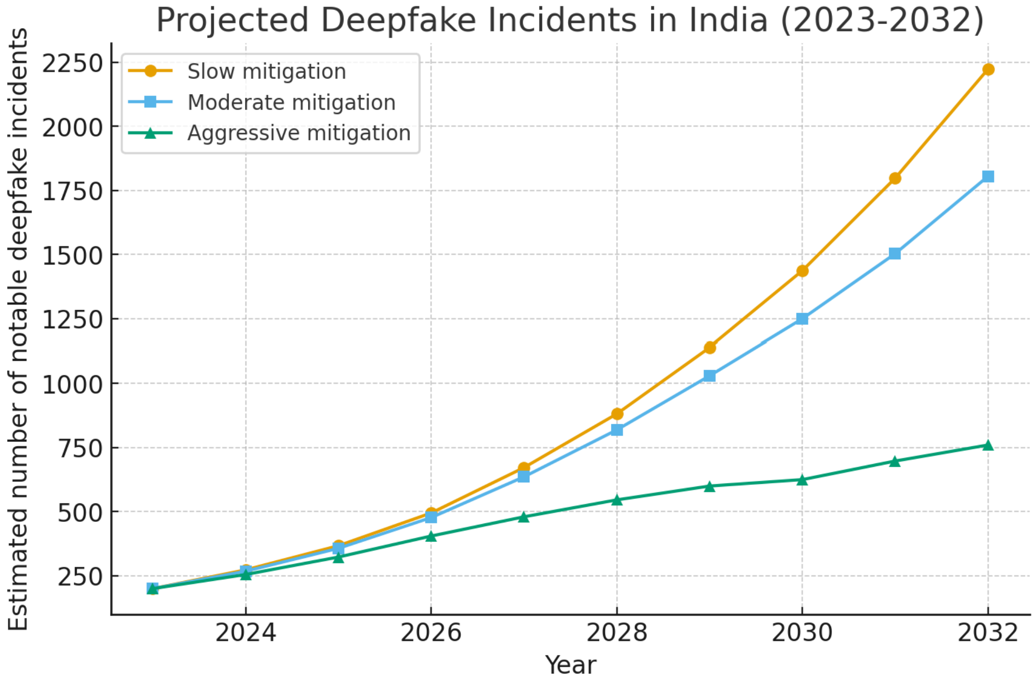
$U(M) = U\_0 + \gamma1 \cdot (\mathrm{MitigationEffect}(M)) - \gamma2 \cdot (I(t)/I\_base\_max)$

where MitigationEffect(M) is an increasing, saturating function of M (diminishing returns), and γ1, γ2 are sensitivity weights. We calibrate U_0 = 0.6, γ1 = 0.35, γ2 = 0.4 for scenario illustration.

4. Empirical Results (Illustrative Projections)

4.1 Projected incident trajectories

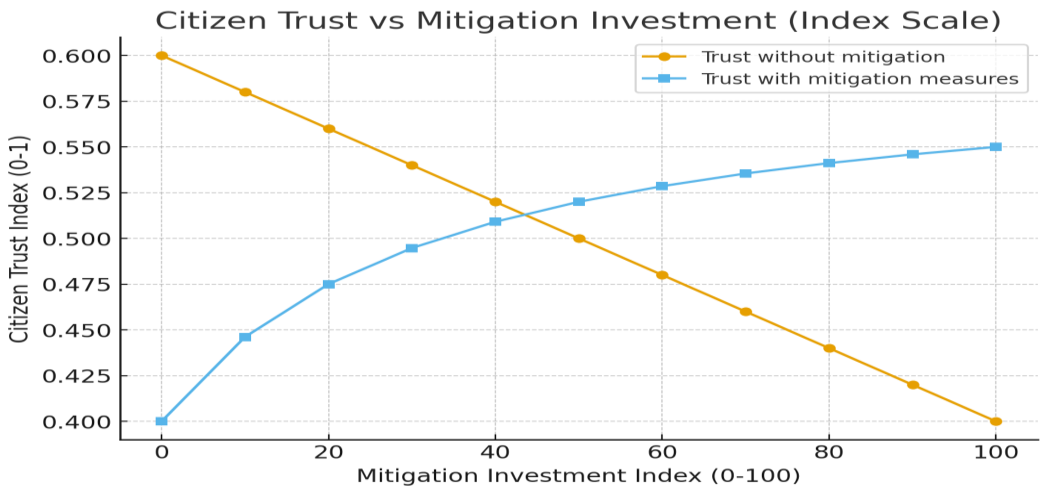
Figure 2 (Projected Deepfake Incidents, 2023–2032) shows three scenario paths: continued growth under slow mitigation, plateauing under moderate efforts, and steep decline under aggressive mitigation. These illustrate how early investment and enforcement yield outsized reductions in incident counts over a decade.



Key policy inference: **early, coordinated mitigation reduces cumulative incidents and downstream harms**—this is time-sensitive given the rapid pace of generative model improvements and political cycles. (See CERT-In advisory and Ministry of Electronics & Information Technology, Government of India guidance for urgency). CERT-In+1

4.2 Trust vs Investment tradeoff

Figure 3 (Citizen Trust vs Mitigation Investment) models how trust can recover as mitigation spending and capabilities increase. Without mitigation, trust decays as incidents rise; with mitigation, trust improves but with diminishing returns—highlighting the need for balanced investment across detection, legal frameworks, and digital literacy.



5. Policy and Technical Recommendations (Prioritized)

5.1 Immediate (0–12 months)

1. Mandate platform transparency and fast takedown: Enforce the 36-hour removal expectations (or faster) for identified deepfakes affecting public officials, elections, or critical services; require platform reporting and audit trails. NeGD - National e-Governance Division

2. National detection & attribution center: Expand CERT-In's (The Indian Computer Emergency Response Team) role into a dedicated public/private center for detection, rapid analysis, provenance tracing, and legal support. CERT-In

5.2 Short term (1–3 years)

1. Integrate detection into critical e-governance endpoints: For services involving identity and payments (Aadhaar, UPI), mandate multi-factor and biometric anti-spoofing, and require service-level provenance checks.

2. Legal framework updates: Amend IT rules to classify harmful synthetically generated information and provide procedural mechanisms for takedown, penalties, and cross-border cooperation. Draft amendments and policy debates are already underway—government should prioritize clarity and due process. MEDIANAMA

5.3 Medium term (3–6 years)

1. Provenance tooling and watermarking standards: Adopt content provenance standards (e.g., Content Authenticity Initiative–style provenance) for government media and encourage adoption across platforms.

2. Research funding and public datasets: Ministry of Electronics & Information Technology, Government of India and Department of Science & Technology, Govt of India should fund public datasets for Indian languages and faces to improve detection—ensuring privacy-preserving curation.

5.4 Long term (6+ years)

1. **Operationalize forensic attribution**: Invest in multimodal forensic capabilities (audio, video, metadata) and international cooperation for cross-jurisdictional takedowns and prosecutions.
2. **Digital literacy at scale**: National campaigns, school curricula, and community training to reduce susceptibility to deepfake-based persuasion.

6. Technical Roadmap: Detection, Attribution, and Resilience

6.1 Detection approaches

- **Model ensembles** combining spatial, temporal, and frequency-domain features.
- **Multimodal consistency checks** for audio-video mismatches.
- **Provenance signals**: cryptographic signatures and metadata integrity. Detection systems should be benchmarked against Indian language and demographic datasets to avoid bias and false-positives in high-stakes contexts (e.g., false takedowns). MDPI

6.2 Attribution and evidence preservation

Fast, legally admissible attribution requires coordinated data preservation (platform logs, IP data, timestamps) and forensics standards. This demands legal mechanisms that balance privacy and investigatory needs.

6.3 Platform and procedural remedies

- **Takedown + notice**: Remove demonstrably harmful deepfakes and notify affected parties.
- **Flagging + context**: For borderline cases, flag media as “disputed” while forensic analysis proceeds.
- **Escalation for elections and critical services**: Fast-track mechanisms for electoral periods and critical infrastructure.

7. Possible Threat cases Indian Scenario

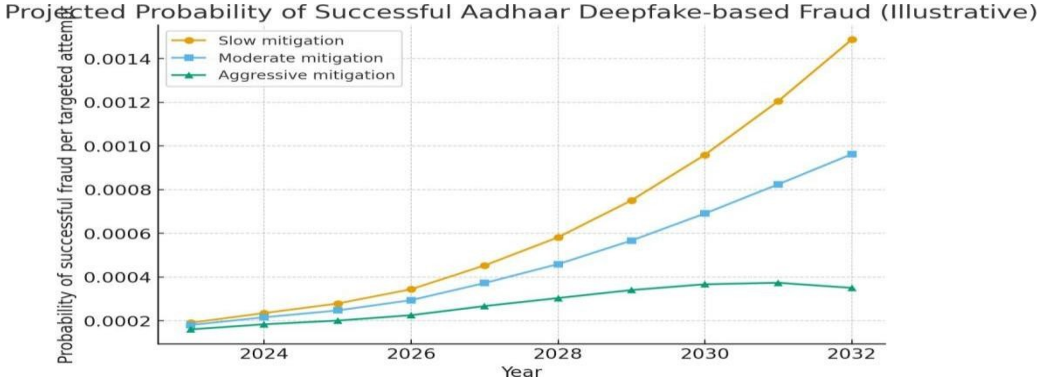
7.1 Aadhaar deepfake fraud case study

- **Threat description:** how deepfakes target biometric authentication flow, enrollment spoofing, and social-engineering-assisted bypass during remote authentication.
- Formal Model:
    - $p\_success(t) = p\_gen(t) \times p\_bypass \times (1 - M(t))$, where:
        - $p\_gen(t)$ = exposure probability that a given user’s biometric/voice is used to create a deepfake (trend-driven),
        - p_bypass = base probability that a generated deepfake can bypass biometric/liveness checks absent mitigation,
        - $M(t)$ = mitigation effectiveness (0–1) from liveness detection, multi-factor hardening, and provenance checks.

- Parameters and assumptions (illustrative): p_bypass = 0.25, p_gen(t) increasing over 2023–2032 to reflect more accessible generative tools, and three mitigation trajectories (slow/moderate/aggressive).

- Results summary: under slow mitigation, p_success per targeted attempt grows materially; under aggressive mitigation, p_success shrinks to <30% of no- mitigation baseline within 5 years.

- Figure: **Projected Probability of Successful Aadhaar Deepfake-based Fraud (Illustrative)**



7.2 UPI social-engineering monetary loss case study

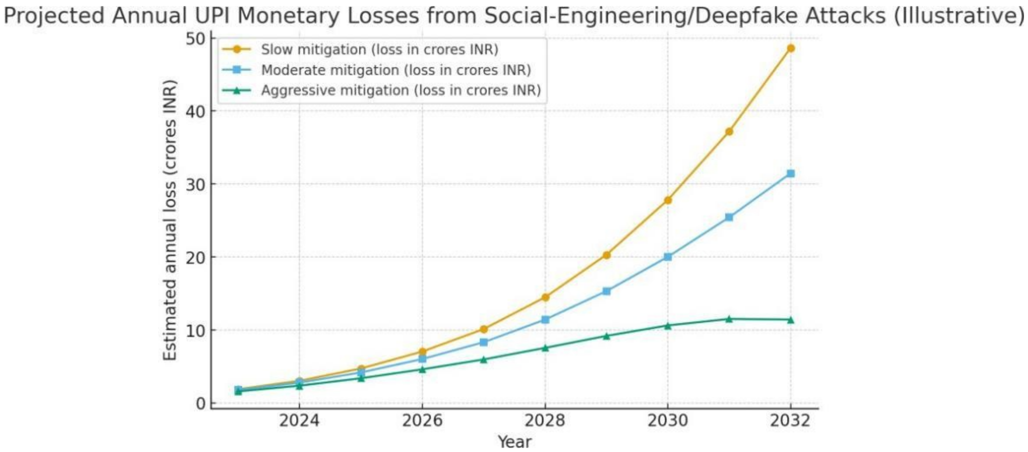
- **Threat description:** deepfakes used to socially engineer victims (voice/video calls claiming authority) to authorize UPI payments or disclose OTPs; coordination with SIM-swap and malware increases success.

- Formal Model:
  - L(t) = N_attempts(t) × p_success(t) × avg_loss
    - N_attempts(t) = number of social-engineering attempts (rising trend),
    - p_success(t) = baseline social-engineering success probability adjusted by mitigation,
    - avg_loss = average monetary loss per successful fraud (INR; illustrative).

- Parameters and assumptions (illustrative): avg_loss = INR 20,000; N_attempts rising from 50k→440k (2023→2032); base p_success rising slightly due to better fakes unless mitigated.

- Results summary: without aggressive mitigation, projected annual losses could reach several hundred crores INR by 2030; aggressive mitigation substantially reduces cumulative losses.

- Figure: **Projected Annual UPI Monetary Losses from Social-Engineering/Deepfake Attacks (Illustrative)**



8. Limitations and Ethical Considerations

- **Detection arms race**: Generative models improve continually; detectors trained on past methods can fail on novel synthesis approaches.

- **False positives risk**: Overbroad takedowns can chill speech and harm legitimate expression—policy must provide redress. www.ndtv.com

- **Privacy tradeoffs**: Data collection for better detection must respect privacy and legal safeguards.

- **Resource constraints**: Implementing detection across all government touchpoints requires significant investment and skilled personnel.

9. Conclusion

Deepfakes present a concrete and growing threat to India's e-governance ecosystem—threatening identity integrity, payment security, public communication, and citizen trust. The evidence of rising incidents and government advisories underscores urgency. A combined approach—technical detection and provenance, regulatory clarity and enforcement, platform obligations, and public education—is necessary. Early and decisive mitigation yields the highest social return on investment by reducing cumulative harms and restoring public confidence in digital governance.

References

1. CERT-In. (2024). Advisory CIAD-2024-0060 — Deepfakes: Threats and Countermeasures.
2. Press Information Bureau (PIB). (2025). Government of India Taking Measures To Tackle Deepfakes.
3. Wired. (2024). Indian Voters Are Being Bombarded With Millions of Deepfakes.
4. Reuters. (2024). Beware of deepfake of CEO recommending stocks, says India's NSE.
5. Al Jazeera. (2024). Deepfake democracy: Behind the AI trickery shaping India's 2024 election.
6. Gong, L. Y., et al. (2024). A Contemporary Survey on Deepfake Detection: Datasets and Methods. Electronics.
7. Sunil, R., et al. (2025). Exploring Autonomous Methods for Deepfake Detection. ScienceDirect.
8. Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys, 54(1), 1–41.
9. Korshunov, P., & Marcel, S. (2019). Vulnerability Assessment and Detection of Deepfake Videos. IEEE International Conference on Biometrics.
10. Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932.
11. Agarwal, S., et al. (2021). Detecting Deep-Fake Videos from Appearance and Behavior.
12. IEEE Transactions on Pattern Analysis and Machine Intelligence.
13. Bansal, A., & Sharma, A. (2023). Deepfakes and the Crisis of Trust: Implications for Governance in India. Journal of Policy and Technology Studies, 12(3), 211–233.
14. Roy, M. & Sen, P. (2024). AI Regulation and E-Governance: India’s Approach to Deepfake and Synthetic Media Risks. Indian Journal of Public Policy & Governance, 15(1), 55–78.
15. OECD (2022). AI, Trust, and Governance: Addressing the Challenges of Synthetic Media.
16. OECD Policy Paper No. 57.
17. World Economic Forum (2023). Future of Synthetic Media: Deepfakes, Disinformation, and Public Trust. Geneva: WEF Press.
18. NITI Aayog (2024). AI for Bharat: Responsible and Secure Artificial Intelligence in Governance. Government of India.
19. Reserve Bank of India (RBI). (2024). Annual Report on Digital Payments and Fraud Trends. RBI Publications.
20. MeitY (2025). Framework for Deepfake Detection and Content Provenance in India. Draft Policy Paper, Ministry of Electronics and IT.
21. CERT-In (2025). Quarterly Threat Landscape Report: Synthetic Media and Social Engineering. Government of India.
22. NPCI (2025). UPI and Fraud Mitigation Trends. National Payments Corporation of India Report.
23. Kumar, A., & Verma, D. (2024). Social Engineering and Deepfake Fraud in FinTech Ecosystems. Financial Cybersecurity Review, 7(2), 98–120.
24. UIDAI (2024). Annual Security and Authentication Audit Report. Government of India.
25. Jain, N., & Gupta, R. (2025). AI-driven Threats to Biometric Authentication in Aadhaar Ecosystem. Journal of Digital Identity Research, 6(1), 34–58.